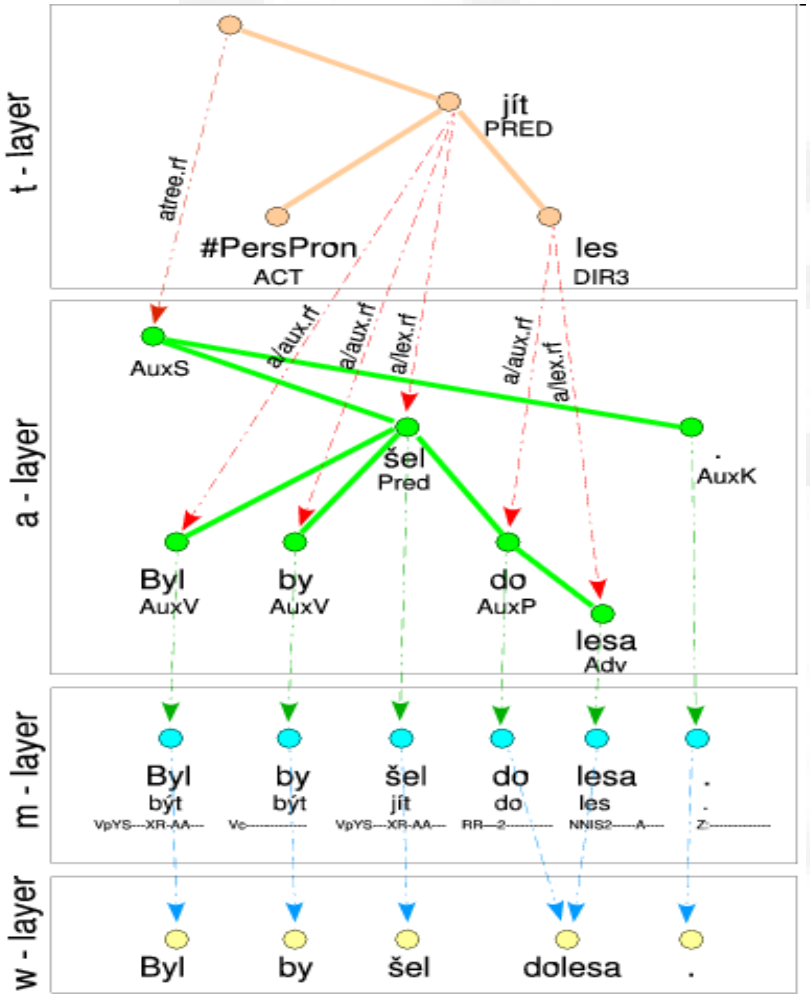
Koncepce roviny diskurzu v popisu jazyka a její začlenění do Pražského závislostního korpusu

Lucie Mladová

mladova@ufal.mff.cuni.cz

Res Informatica, 10. dubna 2009

Roviny popisu přirozeného jazyka v PDT 2.0



Pražský závislostní korpus (PDT):

- 3 roviny popisu jazyka + čistý text (w-layer)
- směrem vzhůru vyšší stupeň zobecnění od formy k významu
- **koncepce nové, „vyšší" roviny – popis textu (promluvy, diskurzu)**
- zachycení textových jednotek a sémantických popř. rétorických vztahů mezi nimi; překročení hranice věty

Koherence textu

- propojení textových jednotek v koherentní, smysluplný sled informací
- zejména mezivětné vztahy
- textové konektory

Nikdy netrávila večery doma. Chodila <u>například</u> na procházky s přáteli. Exemplifikace

Jel do zatáček opatrně. Vždy si nadjížděl. Specifikace

Metoda záleží jenom na vás. <u>Prostě</u> to udělejte podle sebe. Ekvivalence

Syntaktické vs. textové vztahy

- podmínková závislá klauze

  Usmažím palačinky, pokud mi koupíš vajíčka.

x podmínka jako typ textového vztahu:

  Usmažím palačinky. Musíš mi ale nejdřív koupit vajíčka.

- výjimka jako slovesné doplnění

  Nevěnuji se žádnému sportu kromě plavání.

x výjimka jako typ textového vztahu:

  Nevěnuji se žádnému sportu. Jen si občas chodím zaplavat.

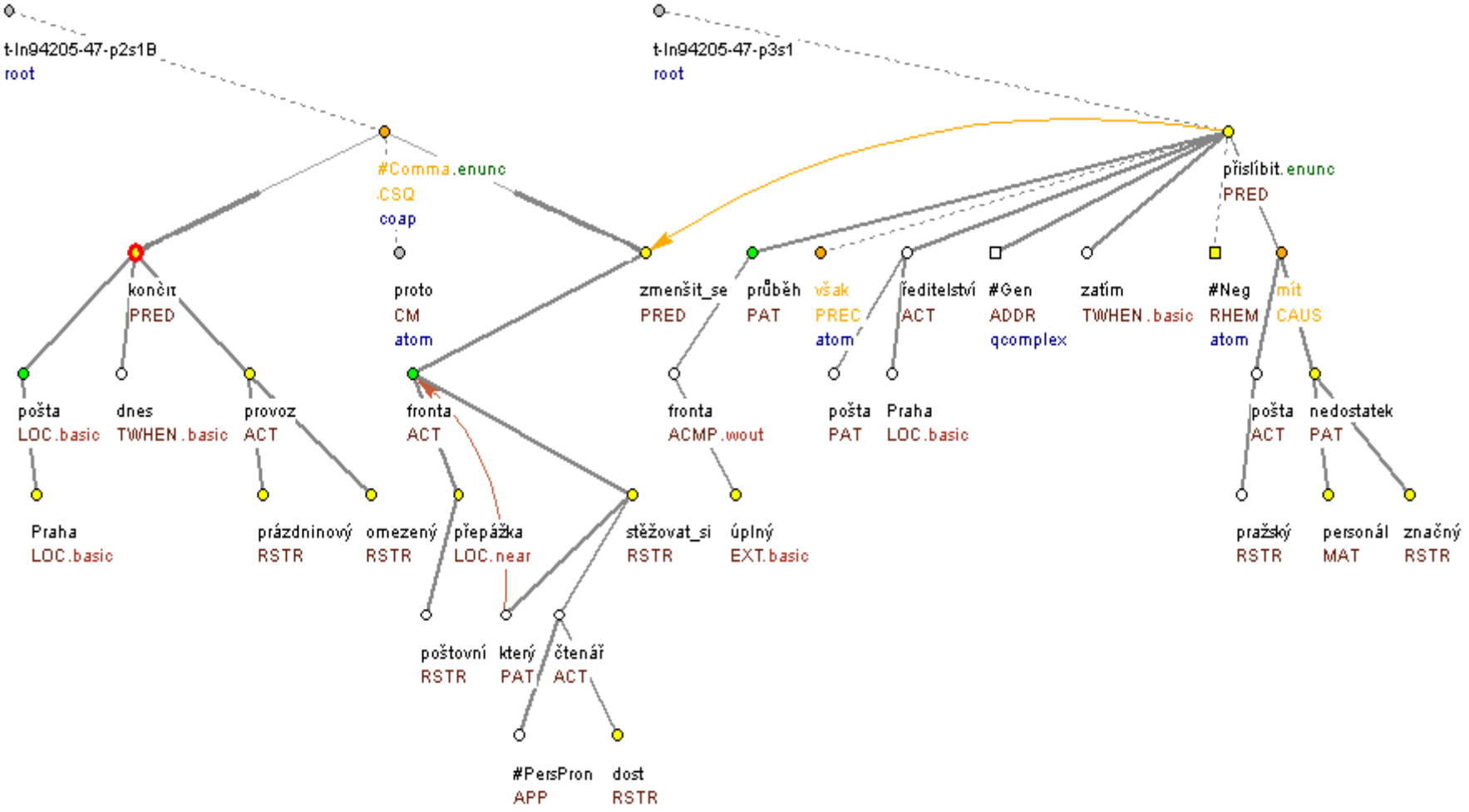
K čemu analýza textových vztahů?

- podklad pro lingvistické analýzy založené na korpusech (první korpus textových vztahů zaměřený na češtinu)
- automatická sumarizace textu
- získávání a odvozování informací
- dialogové systémy
- vztah otázka – odpověď
- automatická anotace dalších korpusů

Zachycení v PDT

- Propojení tektogramatických stromových struktur (vět) šipkami
- Atribut šipky = typ sémantického vztahu mezi spojovanými jednotkami
- Zvýraznění konektorů, jsou-li přítomné
- Zvlášť anotovány anaforické řetězce (Pavel – on – bratr atd.), ve výstupních datech budou ovšem obě tyto anotace společně

Zachycení v PDT



Moje práce

- Teoretické lingvistické podklady pro zpracování této úrovně přirozeného jazyka
- Vypracování anotačního manuálu
- Podíl na úpravě anotačního nástroje (TrEd) pro potřeby anotace diskurzu
- Vedení ručních anotací
- Kontroly a evaluace ručně anotovaných dat

(povahou větší projekt, většina prací probíhá týmově)